

Generalization Bounds for Metric and Similarity Learning

Qiong Cao

Zheng-Chu Guo

Yiming Ying

College of Engineering, Mathematics and Physical Sciences

University of Exeter

Harrison Building, Exeter, EX4 4QF, UK

QC218@EXETER.AC.UK

Z.GUO@EXETER.AC.UK

Y.YING@EXETER.AC.UK

Abstract

Recently, metric learning and similarity learning have attracted a large amount of interest. Many models and optimisation algorithms have been proposed. However, there is relatively little work on the generalization analysis of such methods. In this paper, we derive novel generalization bounds of metric and similarity learning. In particular, we first show that the generalization analysis reduces to the estimation of the Rademacher average over “sums-of-i.i.d.” sample-blocks related to the specific matrix norm. Then, we derive generalization bounds for metric/similarity learning with different matrix-norm regularisers by estimating their specific Rademacher complexities. Our analysis indicates that sparse metric/similarity learning with L^1 -norm regularisation could lead to significantly better bounds than those with Frobenius-norm regularisation. Our novel generalization analysis develops and refines the techniques of U-statistics and Rademacher complexity analysis.

Keywords: Metric learning, Similarity learning, Generalization bound, Rademacher complexity, U-statistics

1. Introduction

The success of many machine learning algorithms (e.g. the nearest neighborhood classification and k-means clustering) depends on the concepts of distance metric and similarity. For instance, k-nearest-neighbor (kNN) classifier depends on a distance function to identify the nearest neighbors for classification; k-means algorithms depend on the pairwise distance measurements between examples for clustering. Kernel methods and information retrieval methods rely on a similarity measure between samples. Many existing studies have been devoted to learning a metric or similarity automatically from data, which is usually referred to as *metric learning* and *similarity learning*, respectively.

Most work in metric learning focuses on learning a (squared) Mahalanobis distance defined, for any $x, t \in \mathbb{R}^d$, by $d_M(x, t) = (x - t)^T M (x - t)$ where M is a positive semi-definite matrix, see e.g. Bar-Hillel et al. (2005); Davis et al. (2007); Globerson and Roweis (2005); Goldberger et al. (2004); Shen et al. (2009); Weinberger and Saul (2008); Xing et al. (2002); Ying et al. (2009); Yang and Jin (2007). Concurrently, the pairwise similarity defined by $s_M(x, t) = x^T M t$ was studied in Chechik et al. (2010); Shalit et al. (2010); Kar and Jain (2011); Maurer (2008). These methods have been successfully applied to various real-world problems including information retrieval and face verification (Chechik et al., 2010; Guillaumin et al., 2009; Hoi et al., 2006; Ying and Li, 2012). Although there are a large

number of studies devoted to supervised metric/similarity learning based on different objective functions, few studies address the generalization analysis of such methods. The recent work (Jin et al., 2009) pioneered the generalization analysis for metric learning using the concept of uniform stability (Bousquet and Elisseeff, 2002). However, this approach only works for the strongly convex norm, e.g. the Frobenius norm, and the offset term is fixed which makes the generalization analysis essentially different.

In this paper, we develop a novel approach for generalization analysis of metric learning and similarity learning which can deal with general matrix regularisation terms including Frobenius norm (Jin et al., 2009), sparse L^1 -norm (Rosales and Fung, 2006), mixed $(2, 1)$ -norm (Ying et al., 2009) and trace-norm (Ying et al., 2009; Shen et al., 2009). In particular, we first show that the generalization analysis for metric/similarity learning reduces to the estimation of the Rademacher average over “sums-of-i.i.d.” sample-blocks related to the specific matrix norm, which we refer to as the *Rademacher complexity for metric (similarity) learning*. Then, we show how to estimate the Rademacher complexities with different matrix regularisers. Our analysis indicates that sparse metric/similarity learning with L^1 -norm regularisation could lead to significantly better generalization bounds than that with Frobenius norm regularisation, especially when the dimension of the input data is high. This is nicely consistent with the rationale that sparse methods are more effective for high-dimensional data analysis. Our novel generalization analysis develops and extends Rademacher complexity analysis (Bartlett and Mendelson, 2002; Koltchinskii and Panchenko, 2002) to the setting of metric/similarity learning by using techniques of U-statistics (Cl  mencon et al., 2008; Pe  a and Gin  , 1999).

The paper is organized as follows. The next section reviews the models of metric/similarity learning. Section 3 establishes the main theorems. In Section 4, we derive and discuss generalization bounds for metric/similarity learning with various matrix-norm regularisation terms. Section 5 concludes the paper.

Notation: Let $\mathbb{N}_n = \{1, 2, \dots, n\}$ for any $n \in \mathbb{N}$. For any $X, Y \in \mathbb{R}^{d \times n}$, $\langle X, Y \rangle = \text{Tr}(X^\top Y)$ where $\text{Tr}(\cdot)$ denotes the trace of a matrix. The space of symmetric d times d matrices will be denoted by \mathbb{S}^d . We equip \mathbb{S}_+^d with a general matrix norm $\|\cdot\|$; it can be a Frobenius norm, trace-norm and mixed norm. Its associated dual norm is denoted, for any $M \in \mathbb{S}^d$, by $\|M\|_* = \sup\{\langle X, M \rangle : X \in \mathbb{S}^d, \|X\| \leq 1\}$. The Frobenius norm on matrices or vector is always denoted by $\|\cdot\|_F$. The cone of positive semi-definite matrices is denoted by \mathbb{S}_+^d . Later on we use the conventional notation that $X_{ij} = (x_i - x_j)(x_i - x_j)^\top$ and $\tilde{X}_{ij} = x_i x_j^\top$.

2. Metric/Similarity Learning Formulation

In our learning setting, we have a input space $\mathcal{X} \subseteq \mathbb{R}^d$ and output (labels) space \mathcal{Y} . Denote $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and suppose $\mathbf{z} := \{z_i = (x_i, y_i) \in \mathcal{Z} : i \in \mathbb{N}_n\}$ an i.i.d. training set according to an unknown distribution ρ on \mathcal{Z} . Denote the $d \times n$ input data matrix by $\mathbf{X} = (x_i : i \in \mathbb{N}_n)$ and the $d \times d$ distance matrix by $M = (M_{\ell k})_{\ell, k \in \mathbb{N}_d}$. Then, the (pseudo-) distance between x_i and x_j is measured by

$$d_M(x_i, x_j) = (x_i - x_j)^\top M (x_i - x_j).$$

The goal of metric learning is to identify a distance function $d_M(x_i, x_j)$ such that it yields a small value for a similar pair and a large value for a dissimilar pair. The bilinear similarity

function is defined by

$$s_M(x_i, x_j) = x_i^\top M x_j.$$

Similarly, the target of similarity learning is to learn $M \in \mathbb{S}^d$ such that it reports a large similarity value for a similar pair and a small similarity value for a dissimilar pair. It is worth of pointing out that we do not require the positive semi-definiteness of the matrix M throughout this paper. However, we do assume M to be symmetric, since this will guarantee the distance (similarity) between x_i and x_j ($d_M(x_i, x_j)$) is equivalent to that between x_j and x_i ($d_M(x_j, x_i)$).

There are two main terms in the metric/similarity learning model: *empirical error* and *matrix regularisation term*. The empirical error function is to employ the similarity and dissimilarity information provided by the label information and the appropriate matrix regularisation term is to avoid overfitting and improve generalization performance.

For any pair of samples (x_i, x_j) , let $r(y_i, y_j) = 1$ if $y_i = y_j$ otherwise $r(y_i, y_j) = -1$. It is expected that there exists an offset term $b \in \mathbb{R}$ such that $d_M(x_i, x_j) \leq b$ for $r(y_i, y_j) = 1$ and $d_M(x_i, x_j) > b$ otherwise. This naturally leads to the empirical error (Jin et al., 2009) defined by

$$\mathcal{E}_{\mathbf{z}}(M, b) := \frac{1}{n(n-1)} \sum_{i,j \in \mathbb{N}_n, i \neq j} I[r(y_i, y_j)(d_M(x_i, x_j) - b) > 0]$$

where the indicator function $I[x]$ equal 1 if x is true and zero otherwise.

Due to the indicator function, the above empirical error is non-differentiable and non-convex which is difficult to do optimisation. A usual way to overcome this shortcoming is to upper-bound it with a differentiable and convex loss function. For instance, we can use the hinge loss to upper-bound the indicator function which leads to the following empirical error:

$$\mathcal{E}_{\mathbf{z}}(M, b) := \frac{1}{n(n-1)} \sum_{i,j \in \mathbb{N}_n, i \neq j} [1 + r(y_i, y_j)(d_M(x_i, x_j) - b)]_+ \quad (1)$$

In order to avoid overfitting, we need to enforce a regularisation term denoted by $\|M\|$, which will restrict the complexity of the distance matrix. We emphasize here $\|\cdot\|$ denotes a general matrix norm in the linear space \mathbb{S}^d . Putting the regularisation term and the empirical error term together yields the following metric learning model:

$$(M_{\mathbf{z}}, b_{\mathbf{z}}) = \arg \min_{M \in \mathbb{S}^d, b \in \mathbb{R}} \{ \mathcal{E}_{\mathbf{z}}(M, b) + \lambda \|M\|^2 \}, \quad (2)$$

where $\lambda > 0$ is a trade-off parameter.

Different regularisation terms lead to different metric learning formulations. For instance, the Frobenius norm $\|M\|_F$ is used in Jin et al. (2009). To favor the element-sparsity, Rosales and Fung (2006) introduced the L^1 -norm regularisation $\|M\| = \sum_{\ell, k \in \mathbb{N}_d} |M_{\ell k}|$. Ying et al. (2009) proposed the mixed $(2, 1)$ -norm $\|M\| = \sum_{\ell \in \mathbb{N}_d} (\sum_{k \in \mathbb{N}_d} |M_{\ell k}|^2)^{\frac{1}{2}}$ to encourage the column-wise sparsity of the distance matrix. The trace-norm regularisation $\|M\| = \sum_{\ell} \sigma_{\ell}(M)$ was also considered by Ying et al. (2009); Shen et al. (2009). Here, $\{\sigma_{\ell} : \ell \in \mathbb{N}_d\}$ denote the singular values of a matrix $M \in \mathbb{S}^d$. Since M is symmetric, the singular values of M are identical to the absolute values of its eigenvalues.

In analogy to the formulation of metric learning, we consider the following empirical error for similarity learning (Maurer, 2008; Chechik et al., 2010):

$$\tilde{\mathcal{E}}_{\mathbf{z}}(M, b) := \frac{1}{n(n-1)} \sum_{i,j \in \mathbb{N}_n, i \neq j} [1 - r(y_i, y_j)(s_M(x_i, x_j) - b)]_+. \quad (3)$$

This leads to the regularised formulation for similarity learning defined as follows:

$$(\tilde{M}_{\mathbf{z}}, \tilde{b}_{\mathbf{z}}) = \arg \min_{M \in \mathbb{S}^d, b \in \mathbb{R}} \{ \tilde{\mathcal{E}}_{\mathbf{z}}(M, b) + \lambda \|M\|^2 \}. \quad (4)$$

Maurer (2008) used the Frobenius-norm regularisation for similarity learning. The trace-norm regularisation has been used by Shalit et al. (2010) to encourage a low-rank similarity matrix M .

3. Statistical Generalization Analysis

In this section, we mainly give a detailed proof of generalization bounds for metric and similarity learning. In particular, we develop a novel line of generalization analysis for metric and similarity learning with general matrix regularisation terms. The key observation is that the empirical data term $\mathcal{E}_{\mathbf{z}}(M, b)$ for metric learning is a modification of U-statistics and it is expected to converge to its expected form defined by

$$\mathcal{E}(M, b) = \iint (1 + r(y, y')(d_M(x, x') - b))_+ d\rho(x, y) d\rho(x', y'). \quad (5)$$

The empirical term $\tilde{\mathcal{E}}_{\mathbf{z}}(M, b)$ for similarity learning is expected to converge to

$$\tilde{\mathcal{E}}(M, b) = \iint (1 - r(y, y')(s_M(x, x') - b))_+ d\rho(x, y) d\rho(x', y'). \quad (6)$$

The target of generalization analysis is to bound the true error $\mathcal{E}(\mathcal{M}_{\mathbf{z}}, b_{\mathbf{z}})$ by the empirical error $\mathcal{E}_{\mathbf{z}}(M_{\mathbf{z}}, b_{\mathbf{z}})$ for metric learning and $\tilde{\mathcal{E}}(\tilde{M}_{\mathbf{z}}, \tilde{b}_{\mathbf{z}})$ by the empirical error $\tilde{\mathcal{E}}_{\mathbf{z}}(\tilde{M}_{\mathbf{z}}, \tilde{b}_{\mathbf{z}})$ for similarity learning.

In the sequel, we provide a detailed proof for generalization bounds of metric learning. Since the proof for similarity learning is exactly the same as that for metric learning, we only mention the results followed with some brief comments.

3.1. Bounding the Solutions

By the definition of $(M_{\mathbf{z}}, b_{\mathbf{z}})$, we know that

$$\mathcal{E}_{\mathbf{z}}(M_{\mathbf{z}}, b_{\mathbf{z}}) + \lambda \|M_{\mathbf{z}}\|^2 \leq \mathcal{E}_{\mathbf{z}}(0, 0) + \lambda \|0\| = 1$$

which implies that

$$\|M_{\mathbf{z}}\| \leq \frac{1}{\sqrt{\lambda}}. \quad (7)$$

Now we turn our attention to deriving the bound of the offset term $b_{\mathbf{z}}$ by modifying the techniques in Chen et al. (2004) which was originally developed to estimate the offset term of the soft-margin SVM.

Lemma 1 For any samples \mathbf{z} and $\lambda > 0$, let $(M_{\mathbf{z}}, b_{\mathbf{z}})$ be a minimizer of problem (2). Then, it satisfies that

$$\min_{i \neq j} [d_{M_{\mathbf{z}}}(x_i, x_j) - b_{\mathbf{z}}] \leq 1, \quad \max_{i \neq j} [d_{M_{\mathbf{z}}}(x_i, x_j) - b_{\mathbf{z}}] \geq 1. \quad (8)$$

Hence, there holds

$$|b_{\mathbf{z}}| \leq 1 + \left(\max_{i \neq j} \|X_{ij}\|_* \right) \|M_{\mathbf{z}}\|. \quad (9)$$

Proof Recall that $X_{ij} = (x_i - x_j)(x_i - x_j)^\top$ and observe, by the definition of the dual norm $\|\cdot\|_*$, that

$$d_M(x_i, x_j) = \langle X_{ij}, M \rangle \leq \|X_{ij}\|_* \|M\|.$$

Using the above observation, estimation (9) follows directly from inequality (8). Hence, it remains to prove the first statement.

To this end, suppose that $r = \min_{i \neq j} [d_{M_{\mathbf{z}}}(x_i, x_j) - b_{\mathbf{z}}] > 1$. Then, $d_{M_{\mathbf{z}}}(x_i, x_j) - (b_{\mathbf{z}} + r - 1) \geq 1$ for any $i \neq j$. Consequently,

$$\begin{aligned} \mathcal{E}_{\mathbf{z}}(M_{\mathbf{z}}, b_{\mathbf{z}} + r - 1) &= \frac{1}{n(n-1)} \sum_{i \neq j, y_i = y_j} (1 + d_{M_{\mathbf{z}}}(x_i, x_j) - b_{\mathbf{z}} - (r - 1))^2 \\ &< \frac{1}{n(n-1)} \sum_{i \neq j, y_i = y_j} (1 + d_{M_{\mathbf{z}}}(x_i, x_j) - b_{\mathbf{z}})^2 \leq \mathcal{E}_{\mathbf{z}}(M_{\mathbf{z}}, b_{\mathbf{z}}). \end{aligned}$$

Hence, $\mathcal{E}_{\mathbf{z}}(M_{\mathbf{z}}, b_{\mathbf{z}} + r - 1) + \lambda \|M_{\mathbf{z}}\| < \mathcal{E}_{\mathbf{z}}(M_{\mathbf{z}}, b_{\mathbf{z}}) + \lambda \|M_{\mathbf{z}}\|$ which contradicts the definition of the minimizer $(M_{\mathbf{z}}, b_{\mathbf{z}})$. Hence, $r = \min_{i \neq j} [d_{M_{\mathbf{z}}}(x_i, x_j) - b_{\mathbf{z}}] \leq 1$. In analogy to the above argument, it can be shown that $\max_{i \neq j} [d_{M_{\mathbf{z}}}(x_i, x_j) - b_{\mathbf{z}}] \geq 1$ which completes the proof of the lemma. \blacksquare

Denote

$$\mathcal{F} = \left\{ (M, b) : \|M\| \leq 1/\sqrt{\lambda}, \quad |b| \leq 1 + X_* \|M\| \right\}, \quad (10)$$

where

$$X_* = \max_{x, x' \in \mathcal{X}} \|(x - x')(x - x')^\top\|_*.$$

From the above lemma, for any samples \mathbf{z} we can easily see that the solution $(M_{\mathbf{z}}, b_{\mathbf{z}})$ of (2) belongs to the bounded set $\mathcal{F} \subseteq \mathbb{S}^d \times \mathbb{R}$.

3.2. Generalization Bounds

Before stating the generalization bounds, we introduce some notations. For any $z = (x, y)$, $z' = (x', y') \in \mathcal{Z}$, let $\Phi_{M,b}(z, z') = (1 + r(y, y')(d_M(x, x') - b))_+$. Hence, for any $(M, b) \in \mathcal{F}$,

$$\sup_{z, z'} \sup_{(M, b) \in \mathcal{F}} \Phi_{M,b}(z, z') \leq B_\lambda := 2(1 + X_*/\sqrt{\lambda}). \quad (11)$$

Let $\lfloor \frac{n}{2} \rfloor$ denote the largest integer less than $\frac{n}{2}$ and recall the definition that $X_{ij} = (x_i - x_j)(x_i - x_j)^\top$. We now define Rademacher average over sums-of-i.i.d. sample-blocks related to the dual matrix norm $\|\cdot\|_*$ by

$$\hat{R}_n = \frac{1}{\lfloor \frac{n}{2} \rfloor} \mathbb{E}_\sigma \left\| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i X_{i(\lfloor \frac{n}{2} \rfloor + i)} \right\|_*, \quad (12)$$

and its expectation is denoted by $R_n = \mathbb{E}_{\mathbf{z}}[\widehat{R}_n]$. Our main theorem below shows that the generalization bounds for metric learning critically depend on the quantity of R_n . For this reason, we refer to R_n as the *Radmemcher complexity for metric learning*. It is worth mentioning that metric learning formulation (2) depends on the norm $\|\cdot\|$ of the linear space \mathbb{S}^d and the Rademacher complexity R_n is related to its dual norm $\|\cdot\|_*$.

Theorem 2 *Let $(M_{\mathbf{z}}, b_{\mathbf{z}})$ be the solution of formulation (2). Then, for any $0 < \delta < 1$, with probability $1 - \delta$ we have that*

$$\begin{aligned} \mathcal{E}(M_{\mathbf{z}}, b_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(M_{\mathbf{z}}, b_{\mathbf{z}}) &\leq \sup_{(M,b) \in \mathcal{F}} [\mathcal{E}(M, b) - \mathcal{E}_{\mathbf{z}}(M, b)] \\ &\leq \frac{4R_n}{\sqrt{\lambda}} + \frac{4(3+2X_*/\sqrt{\lambda})}{\sqrt{n}} + 2(1 + X_*/\sqrt{\lambda}) \left(\frac{2\ln(\frac{1}{\delta})}{n} \right)^{\frac{1}{2}}. \end{aligned} \quad (13)$$

Proof The proof of the theorem can be divided into three steps as follows.

Step 1: Let $\mathbb{E}_{\mathbf{z}}$ denote the expectation with respect to samples \mathbf{z} . Observe that $\mathcal{E}(M_{\mathbf{z}}, b_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(M_{\mathbf{z}}, b_{\mathbf{z}}) \leq \sup_{(M,b) \in \mathcal{F}} [\mathcal{E}(M, b) - \mathcal{E}_{\mathbf{z}}(M, b)]$. For any $z = (z_1, \dots, z_{k-1}, z_k, z_{k+1}, \dots, z_n)$ and $z' = (z_1, \dots, z_{k-1}, z'_k, z_{k+1}, \dots, z_n)$ we know from inequality (11) that

$$\begin{aligned} &\left| \sup_{(M,b) \in \mathcal{F}} [\mathcal{E}(M, b) - \mathcal{E}_{\mathbf{z}}(M, b)] - \sup_{(M,b) \in \mathcal{F}} [\mathcal{E}(M, b) - \mathcal{E}_{\mathbf{z}'}(M, b)] \right| \\ &\leq \sup_{(M,b) \in \mathcal{F}} |\mathcal{E}_{\mathbf{z}}(M, b) - \mathcal{E}_{\mathbf{z}'}(M, b)| \\ &= \frac{1}{n(n-1)} \sup_{(M,b) \in \mathcal{F}} \sum_{j \in \mathbb{N}_n, j \neq k} |\Phi_{M,b}(z_k, z_j) - \Phi_{M,b}(z'_k, z_j)| \\ &\leq \frac{1}{n(n-1)} \sup_{(M,b) \in \mathcal{F}} \sum_{j \in \mathbb{N}_n, j \neq k} |\Phi_{M,b}(z_k, z_j)| + |\Phi_{M,b}(z'_k, z_j)| \\ &\leq 4(1 + X_*/\sqrt{\lambda})/n. \end{aligned}$$

Applying McDiarmid's inequality (McDiarmid, 1989) (see Lemma 5 in the Appendix) to the term $\sup_{(M,b) \in \mathcal{F}} [\mathcal{E}(M, b) - \mathcal{E}_{\mathbf{z}}(M, b)]$, with probability $1 - \delta$ there holds

$$\begin{aligned} \sup_{(M,b) \in \mathcal{F}} [\mathcal{E}(M, b) - \mathcal{E}_{\mathbf{z}}(M, b)] &\leq \mathbb{E}_{\mathbf{z}} \sup_{(M,b) \in \mathcal{F}} [\mathcal{E}(M, b) - \mathcal{E}_{\mathbf{z}}(M, b)] \\ &\quad + 2(1 + X_*/\sqrt{\lambda}) \left(\frac{2\ln(\frac{1}{\delta})}{n} \right)^{\frac{1}{2}}. \end{aligned} \quad (14)$$

Now we only need to estimate the first term in the expectation form on the right-hand side of the above equation by symmetrization techniques.

Step 2: To estimate $\mathbb{E}_{\mathbf{z}} \sup_{(M,b) \in \mathcal{F}} [\mathcal{E}(M, b) - \mathcal{E}_{\mathbf{z}}(M, b)]$, applying Lemma 6 with $q_{(M,b)}(z_i, z_j) = \mathcal{E}(M, b) - (1 + r(y_i, y_j)(d_M(x_i, x_j) - b))_+$ implies that

$$\mathbb{E}_{\mathbf{z}} \sup_{(M,b) \in \mathcal{F}} [\mathcal{E}(M, b) - \mathcal{E}_{\mathbf{z}}(M, b)] \leq \mathbb{E}_{\mathbf{z}} \sup_{(M,b) \in \mathcal{F}} [\mathcal{E}(M, b) - \overline{\mathcal{E}}_{\mathbf{z}}(M, b)], \quad (15)$$

where $\bar{\mathcal{E}}_{\mathbf{z}}(M, b) = \frac{1}{\lfloor \frac{n}{2} \rfloor} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \Phi_{M,b}(z_i, z_{\lfloor \frac{n}{2} \rfloor + i})$. Now let $\bar{\mathbf{z}} = \{\bar{z}_1, \bar{z}_2, \dots, \bar{z}_n\}$ be i.i.d. samples which are independent of \mathbf{z} , then

$$\begin{aligned} \mathbb{E}_{\mathbf{z}} \sup_{(M,b) \in \mathcal{F}} [\mathcal{E}(M, b) - \bar{\mathcal{E}}_{\mathbf{z}}(M, b)] &= \mathbb{E}_{\mathbf{z}} \sup_{(M,b) \in \mathcal{F}} [\mathbb{E}_{\bar{\mathbf{z}}} [\bar{\mathcal{E}}_{\bar{\mathbf{z}}}(M, b)] - \bar{\mathcal{E}}_{\mathbf{z}}(M, b)] \\ &\leq \mathbb{E}_{\mathbf{z}, \bar{\mathbf{z}}} \sup_{(M,b) \in \mathcal{F}} [\bar{\mathcal{E}}_{\bar{\mathbf{z}}}(M, b) - \bar{\mathcal{E}}_{\mathbf{z}}(M, b)] \end{aligned} \quad (16)$$

By standard symmetrization techniques (see e.g. [Bartlett and Mendelson \(2002\)](#)), for i.i.d. Rademacher variables $\{\sigma_i \in \{\pm 1\} : i \in \mathbb{N}_{\lfloor \frac{n}{2} \rfloor}\}$, we have that

$$\begin{aligned} &\mathbb{E}_{\mathbf{z}, \bar{\mathbf{z}}} \sup_{(M,b) \in \mathcal{F}} [\bar{\mathcal{E}}_{\bar{\mathbf{z}}}(M, b) - \bar{\mathcal{E}}_{\mathbf{z}}(M, b)] \\ &= \mathbb{E}_{\mathbf{z}, \bar{\mathbf{z}}} \frac{1}{\lfloor \frac{n}{2} \rfloor} \sup_{(M,b) \in \mathcal{F}} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i [\Phi_{M,b}(\bar{z}_i, \bar{z}_{\lfloor \frac{n}{2} \rfloor + i}) - \Phi_{M,b}(z_i, z_{\lfloor \frac{n}{2} \rfloor + i})] \\ &= 2\mathbb{E}_{\mathbf{z}, \sigma} \frac{1}{\lfloor \frac{n}{2} \rfloor} \sup_{(M,b) \in \mathcal{F}} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i \Phi_{M,b}(\bar{z}_i, \bar{z}_{\lfloor \frac{n}{2} \rfloor + i}) \\ &\leq 2\mathbb{E}_{\mathbf{z}, \sigma} \frac{1}{\lfloor \frac{n}{2} \rfloor} \sup_{(M,b) \in \mathcal{F}} \left| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i \Phi_{M,b}(\bar{z}_i, \bar{z}_{\lfloor \frac{n}{2} \rfloor + i}) \right|. \end{aligned} \quad (17)$$

Applying the contraction property of Rademacher averages (see Lemma 7 in the Appendix) with $\Psi_i(t) = (1 + r(y_i, y_{\lfloor \frac{n}{2} \rfloor + i})t)_+ - 1$, we have the following estimation for the last term on the righthand side of the above inequality:

$$\begin{aligned} &\mathbb{E}_{\sigma} \frac{1}{\lfloor \frac{n}{2} \rfloor} \sup_{(M,b) \in \mathcal{F}} \left| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i \Phi_{M,b}(\bar{z}_i, \bar{z}_{\lfloor \frac{n}{2} \rfloor + i}) \right| \\ &\leq \mathbb{E}_{\sigma} \frac{1}{\lfloor \frac{n}{2} \rfloor} \sup_{(M,b) \in \mathcal{F}} \left| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i (\Phi_{M,b}(\bar{z}_i, \bar{z}_{\lfloor \frac{n}{2} \rfloor + i}) - 1) \right| + \frac{1}{\lfloor \frac{n}{2} \rfloor} \mathbb{E}_{\sigma} \left| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i \right| \\ &\leq \frac{2}{\lfloor \frac{n}{2} \rfloor} \mathbb{E}_{\sigma} \sup_{(M,b) \in \mathcal{F}} \left| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i (d_M(x_i, x_{\lfloor \frac{n}{2} \rfloor + i}) - b) \right| + \frac{1}{\lfloor \frac{n}{2} \rfloor} \mathbb{E}_{\sigma} \left| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i \right| \\ &\leq \frac{2}{\lfloor \frac{n}{2} \rfloor} \mathbb{E}_{\sigma} \sup_{\|M\| \leq \frac{1}{\sqrt{\lambda}}} \left| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i d_M(x_i, x_{\lfloor \frac{n}{2} \rfloor + i}) \right| + \frac{(3 + 2X_*/\sqrt{\lambda})}{\lfloor \frac{n}{2} \rfloor} \mathbb{E}_{\sigma} \left| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i \right| \end{aligned} \quad (18)$$

Step 3 : It remains to estimate the terms on the righthand side of inequality (18). To this end, observe that

$$\mathbb{E}_{\sigma} \left| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i \right| \leq \left(\mathbb{E}_{\sigma} \left| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i \right|^2 \right)^{\frac{1}{2}} \leq \sqrt{\lfloor \frac{n}{2} \rfloor}.$$

Moreover,

$$\begin{aligned} \mathbb{E}_\sigma \sup_{\|M\|_F \leq \frac{1}{\sqrt{\lambda}}} \left| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i d_M(x_i, x_{\lfloor \frac{n}{2} \rfloor + i}) \right| &\leq \mathbb{E}_\sigma \sup_{\|M\| \leq \frac{1}{\sqrt{\lambda}}} \left| \left\langle \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i (x_i - x_{\lfloor \frac{n}{2} \rfloor + i})(x_i - x_{\lfloor \frac{n}{2} \rfloor + i})^\top, M \right\rangle \right| \\ &= \frac{1}{\sqrt{\lambda}} \mathbb{E}_\sigma \left\| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i X_{i(\lfloor \frac{n}{2} \rfloor + i)} \right\|_*. \end{aligned}$$

Putting the above estimations and inequalities (17), (18) together yields that

$$\mathbb{E}_{\mathbf{z}, \bar{\mathbf{z}}} \sup_{(M, b) \in \mathcal{F}} \left[\bar{\mathcal{E}}_{\bar{\mathbf{z}}}(M, b) - \bar{\mathcal{E}}_{\mathbf{z}}(M, b) \right] \leq \frac{2(3 + 2X_*/\sqrt{\lambda})}{\sqrt{\lfloor \frac{n}{2} \rfloor}} + \frac{4R_n}{\sqrt{\lambda}} \leq \frac{4(3 + X_*/\sqrt{\lambda})}{\sqrt{n}} + \frac{2R_n}{\sqrt{\lambda}}.$$

Consequently, combining this with inequalities (15), (16) implies that

$$\mathbb{E}_{\mathbf{z}} \sup_{(M, b) \in \mathcal{F}} \left[\mathcal{E}(M, b) - \mathcal{E}_{\mathbf{z}}(M, b) \right] \leq \frac{4(3 + 2X_*/\sqrt{\lambda})}{\sqrt{n}} + \frac{4R_n}{\sqrt{\lambda}}.$$

Putting this estimation with (14) completes the proof the theorem. \blacksquare

In the setting of similarity learning, X_* and R_n are replaced by

$$\tilde{X}_* = \sup_{x, t \in \mathcal{X}} \|xt^\top\|_* \quad \text{and} \quad \tilde{R}_n = \frac{1}{\lfloor \frac{n}{2} \rfloor} \mathbb{E}_{\mathbf{z}} \mathbb{E}_\sigma \left\| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i \tilde{X}_{i(\lfloor \frac{n}{2} \rfloor + i)} \right\|_*, \quad (19)$$

where $\tilde{X}_{i(\lfloor \frac{n}{2} \rfloor + i)} = x_i x_{\lfloor \frac{n}{2} \rfloor + i}^\top$. Let $\tilde{\mathcal{F}} = \left\{ (M, b) : \|M\| \leq 1/\sqrt{\lambda}, |b| \leq 1 + \tilde{X}_* \|M\| \right\}$. Using the exactly same argument as above, we can prove the following bound for similarity learning formulation (4).

Theorem 3 *Let $(\tilde{M}_{\mathbf{z}}, \tilde{b}_{\mathbf{z}})$ be the solution of formulation (4). Then, for any $0 < \delta < 1$, with probability $1 - \delta$ we have that*

$$\begin{aligned} \tilde{\mathcal{E}}(\tilde{M}_{\mathbf{z}}, \tilde{b}_{\mathbf{z}}) - \tilde{\mathcal{E}}_{\mathbf{z}}(\tilde{M}_{\mathbf{z}}, \tilde{b}_{\mathbf{z}}) &\leq \sup_{(M, b) \in \tilde{\mathcal{F}}} \left[\tilde{\mathcal{E}}(M, b) - \tilde{\mathcal{E}}_{\mathbf{z}}(M, b) \right] \\ &\leq \frac{4\tilde{R}_n}{\sqrt{\lambda}} + \frac{4(3 + 2\tilde{X}_*/\sqrt{\lambda})}{\sqrt{n}} + 2(1 + \tilde{X}_*/\sqrt{\lambda}) \left(\frac{2 \ln(\frac{1}{\delta})}{n} \right)^{\frac{1}{2}}. \end{aligned} \quad (20)$$

4. Estimation of R_n and Discussion

From Theorem 2, we need to estimate the Rademacher average for metric learning, i.e. R_n , and the quantity X_* for different matrix regularisation terms. Without loss of generality, we only focus on popular matrix norms such as the Frobenius norm (Jin et al., 2009), L^1 -norm (Rosales and Fung, 2006), trace-norm (Ying et al., 2009; Shen et al., 2009) and mixed $(2, 1)$ -norm (Ying et al., 2009).

Example 1 (Frobenius norm) Let the matrix norm be the Frobenius norm i.e. $\|M\| = \|M\|_F$, then the quantity $X_* = \sup_{x, x' \in \mathcal{X}} \|x - x'\|_F^2$ and the Rademacher complexity is estimated as follows:

$$R_n \leq \frac{2X_*}{\sqrt{n}} = \frac{2 \sup_{x, x' \in \mathcal{X}} \|x - x'\|_F^2}{\sqrt{n}}.$$

Let $(M_{\mathbf{z}}, b_{\mathbf{z}})$ be a solution of formulation (2) with Frobenius norm regularisation. For any $0 < \delta < 1$, with probability $1 - \delta$ there holds

$$\begin{aligned} \mathcal{E}(M_{\mathbf{z}}, b_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(M_{\mathbf{z}}, b_{\mathbf{z}}) &\leq 2 \left(1 + \frac{\sup_{x, x' \in \mathcal{X}} \|x - x'\|_F^2}{\sqrt{\lambda}} \right) \sqrt{\frac{2 \ln\left(\frac{1}{\delta}\right)}{n}} \\ &\quad + \frac{16 \sup_{x, x' \in \mathcal{X}} \|x - x'\|_F^2}{\sqrt{n\lambda}} + \frac{12}{\sqrt{n}}. \end{aligned} \quad (21)$$

Proof Note that the dual norm of the Frobenius norm is itself. The estimation of X_* is straightforward. The Rademacher complexity R_n is estimated as follows:

$$\begin{aligned} R_n &= \frac{1}{\lfloor \frac{n}{2} \rfloor} \mathbb{E} \left(\sum_{i,j=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i \sigma_j \langle x_i - x_{\lfloor \frac{n}{2} \rfloor + i}, x_j - x_{\lfloor \frac{n}{2} \rfloor + j} \rangle^2 \right)^{\frac{1}{2}} \\ &\leq \frac{1}{\lfloor \frac{n}{2} \rfloor} \mathbb{E}_{\sigma} \left(\sum_{i,j=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i \sigma_j \langle x_i - x_{\lfloor \frac{n}{2} \rfloor + i}, x_j - x_{\lfloor \frac{n}{2} \rfloor + j} \rangle^2 \right)^{\frac{1}{2}} \\ &= \frac{1}{\lfloor \frac{n}{2} \rfloor} \mathbb{E}_{\mathbf{z}} \left(\sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \|x_i - x_{\lfloor \frac{n}{2} \rfloor + i}\|_F^4 \right)^{\frac{1}{2}} \\ &\leq X_* / \sqrt{\lfloor \frac{n}{2} \rfloor} \leq \frac{2X_*}{\sqrt{n}}. \end{aligned}$$

Putting this estimation back into equation (13) completes the proof of Example 1. \blacksquare

Other popular matrix norms for metric learning are the L^1 -norm, trace-norm and mixed $(2, 1)$ -norm. The dual norms are respectively L^∞ -norm, spectral norm (i.e. the maximum of singular values) and mixed $(2, \infty)$ -norm. All these dual norms mentioned above are less than the Frobenius norm. Hence, the following estimation always holds true for all the norms mentioned above:

$$X_* \leq \sup_{x, x' \in \mathcal{X}} \|x - x'\|_F^2, \quad \text{and} \quad R_n \leq \frac{2 \sup_{x, x' \in \mathcal{X}} \|x - x'\|_F^2}{\sqrt{n}}.$$

Consequently, the generalization bound (21) holds true for metric learning formulation (2) with L^1 -norm, or trace-norm or mixed $(2, 1)$ -norm regularisation. However, in some cases, the above upper-bounds are too conservative. For instance, in the following examples we can show that more refined estimation of R_n can be obtained by applying the Khinchin inequalities for Rademacher averages (Peña and Giné, 1999).

Example 2 (Sparse L^1 -norm) Let the matrix norm be the L^1 -norm i.e. $\|M\| = \sum_{\ell, k \in \mathbb{N}_d} |M_{\ell k}|$. Then, $X_* = \sup_{x, x' \in \mathcal{X}} \|x - x'\|_\infty^2$ and

$$R_n \leq 4 \sup_{x, x' \in \mathcal{X}} \|x - x'\|_\infty^2 \sqrt{\frac{e \log d}{n}}.$$

Let $(M_{\mathbf{z}}, b_{\mathbf{z}})$ be a solution of formulation (2) with L^1 -norm regularisation. For any $0 < \delta < 1$, with probability $1 - \delta$ there holds

$$\begin{aligned} \mathcal{E}(M_{\mathbf{z}}, b_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(M_{\mathbf{z}}, b_{\mathbf{z}}) &\leq 2 \left(1 + \frac{\sup_{x, x' \in \mathcal{X}} \|x - x'\|_{\infty}^2}{\sqrt{\lambda}} \right) \sqrt{\frac{2 \ln(\frac{1}{\delta})}{n}} \\ &\quad + \frac{8 \sup_{x, x' \in \mathcal{X}} \|x - x'\|_{\infty}^2 (1 + 2\sqrt{e \log d})}{\sqrt{n\lambda}} + \frac{12}{\sqrt{n}}. \end{aligned} \quad (22)$$

Proof The dual norm of the L^1 -norm is the L^{∞} -norm. Hence, $X_* = \sup_{x, x' \in \mathcal{X}} \|x - x'\|_{\infty}^2$. To estimate R_n , we observe, for any $1 < q < \infty$, that

$$\begin{aligned} R_n &= \frac{1}{\lfloor \frac{n}{2} \rfloor} \mathbb{E}_{\mathbf{z}} \mathbb{E}_{\sigma} \left\| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i X_{i(\lfloor \frac{n}{2} \rfloor + i)} \right\|_{\infty} \leq \frac{1}{\lfloor \frac{n}{2} \rfloor} \mathbb{E}_{\mathbf{z}} \mathbb{E}_{\sigma} \left\| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i X_{i(\lfloor \frac{n}{2} \rfloor + i)} \right\|_q \\ &:= \frac{1}{\lfloor \frac{n}{2} \rfloor} \mathbb{E}_{\mathbf{z}} \mathbb{E}_{\sigma} \left(\sum_{\ell, k \in \mathbb{N}_d} \left| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i (x_i^k - x_{\lfloor \frac{n}{2} \rfloor + i}^k) (x_i^{\ell} - x_{\lfloor \frac{n}{2} \rfloor + i}^{\ell}) \right|^q \right)^{\frac{1}{q}} \\ &\leq \frac{1}{\lfloor \frac{n}{2} \rfloor} \mathbb{E}_{\mathbf{z}} \left(\sum_{\ell, k \in \mathbb{N}_d} \mathbb{E}_{\sigma} \left| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i (x_i^k - x_{\lfloor \frac{n}{2} \rfloor + i}^k) (x_i^{\ell} - x_{\lfloor \frac{n}{2} \rfloor + i}^{\ell}) \right|^q \right)^{\frac{1}{q}} \end{aligned} \quad (23)$$

where x_i^k represents the k -th coordinate element of vector $x_i \in \mathbb{R}^d$. To estimate the term on the right-hand side of inequality (23), we apply the Khinchin-Kahane inequality (See Lemma 8 in the Appendix) with $p = 2 < q < \infty$ yields that

$$\begin{aligned} &\mathbb{E}_{\sigma} \left| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i (x_i^k - x_{\lfloor \frac{n}{2} \rfloor + i}^k) (x_i^{\ell} - x_{\lfloor \frac{n}{2} \rfloor + i}^{\ell}) \right|^q \\ &\leq q^{\frac{q}{2}} \left(\mathbb{E}_{\sigma} \left| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i (x_i^k - x_{\lfloor \frac{n}{2} \rfloor + i}^k) (x_i^{\ell} - x_{\lfloor \frac{n}{2} \rfloor + i}^{\ell}) \right|^2 \right)^{\frac{q}{2}} \\ &= q^{\frac{q}{2}} \left(\sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} (x_i^k - x_{\lfloor \frac{n}{2} \rfloor + i}^k)^2 (x_i^{\ell} - x_{\lfloor \frac{n}{2} \rfloor + i}^{\ell})^2 \right)^{\frac{q}{2}} \leq \max_{x, x' \in \mathcal{X}} \|x - x'\|_{\infty}^{2q} \left(\lfloor \frac{n}{2} \rfloor \right)^{\frac{q}{2}} q^{\frac{q}{2}}. \end{aligned} \quad (24)$$

Putting the above estimation back into (23) and letting $q = 4 \log d$ implies that

$$\begin{aligned} R_n &\leq \max_{x, x' \in \mathcal{X}} \|x - x'\|_{\infty}^2 d^{\frac{2}{q}} \sqrt{q} / \sqrt{\lfloor \frac{n}{2} \rfloor} = 2 \sup_{x, x' \in \mathcal{X}} \|x - x'\|_{\infty}^2 \sqrt{e \log d / \lfloor \frac{n}{2} \rfloor} \\ &\leq 4 \sup_{x, x' \in \mathcal{X}} \|x - x'\|_{\infty}^2 \sqrt{e \log d / n}. \end{aligned}$$

Putting the estimation for X_* and R_n into Theorem 13 yields inequality (22). This completes the proof of Example 2. \blacksquare

Example 3 (Mixed (2, 1)-norm) Consider $\|M\| = \sum_{\ell \in \mathbb{N}_d} \sqrt{\sum_{k \in \mathbb{N}_d} |M_{\ell k}|^2}$. Then, we have $X_* = [\sup_{x, x' \in \mathcal{X}} \|x - x'\|_F] [\sup_{x, x' \in \mathcal{X}} \|x - x'\|_{\infty}]$, and

$$R_n \leq 4 \left[\sup_{x, x' \in \mathcal{X}} \|x - x'\|_{\infty} \right] \left[\sup_{x, x' \in \mathcal{X}} \|x - x'\|_F \right] \sqrt{\frac{e \log d}{n}}.$$

Let $(M_{\mathbf{z}}, b_{\mathbf{z}})$ be a solution of formulation (2) with mixed (2, 1)-norm. For any $0 < \delta < 1$, with probability $1 - \delta$ there holds

$$\begin{aligned} \mathcal{E}(M_{\mathbf{z}}, b_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(M_{\mathbf{z}}, b_{\mathbf{z}}) &\leq 2 \left(1 + \frac{[\sup_{x, x' \in \mathcal{X}} \|x - x'\|_{\infty}] [\sup_{x, x' \in \mathcal{X}} \|x - x'\|_F]}{\sqrt{\lambda}} \right) \sqrt{\frac{2 \ln(\frac{1}{\delta})}{n}} \\ &\quad + \frac{8 [\sup_{x, x' \in \mathcal{X}} \|x - x'\|_{\infty}] [\sup_{x, x' \in \mathcal{X}} \|x - x'\|_F] (1 + 2\sqrt{e \log d})}{\sqrt{n\lambda}} + \frac{12}{\sqrt{n}}. \end{aligned} \quad (25)$$

Proof The estimation of X_* is straightforward and we estimate R_n as follows. For any $q > 1$, there holds

$$\begin{aligned}
 R_n &= \frac{1}{\lfloor \frac{n}{2} \rfloor} \mathbb{E}_{\mathbf{z}} \mathbb{E}_{\sigma} \left\| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i X_{i(\lfloor \frac{n}{2} \rfloor + i)} \right\|_{(2, \infty)} \\
 &= \frac{1}{\lfloor \frac{n}{2} \rfloor} \mathbb{E}_{\mathbf{z}} \mathbb{E}_{\sigma} \sup_{\ell \in \mathbb{N}_d} \left(\sum_{k \in \mathbb{N}_d} \left| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i (x_i^k - x_{\lfloor \frac{n}{2} \rfloor + i}^k) (x_i^\ell - x_{\lfloor \frac{n}{2} \rfloor + i}^\ell) \right|^2 \right)^{\frac{1}{2}} \\
 &\leq \frac{1}{\lfloor \frac{n}{2} \rfloor} \mathbb{E}_{\mathbf{z}} \left(\sum_{k \in \mathbb{N}_d} \mathbb{E}_{\sigma} \sup_{\ell \in \mathbb{N}_d} \left| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i (x_i^k - x_{\lfloor \frac{n}{2} \rfloor + i}^k) (x_i^\ell - x_{\lfloor \frac{n}{2} \rfloor + i}^\ell) \right|^2 \right)^{\frac{1}{2}}.
 \end{aligned} \tag{26}$$

It remains to estimate the terms inside the parenthesis on the right-hand side of the above inequality. To this end, we observe, for any $q' > 1$, that

$$\begin{aligned}
 &\mathbb{E}_{\sigma} \sup_{\ell \in \mathbb{N}_d} \left| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i (x_i^k - x_{\lfloor \frac{n}{2} \rfloor + i}^k) (x_i^\ell - x_{\lfloor \frac{n}{2} \rfloor + i}^\ell) \right|^2 \\
 &\leq \mathbb{E}_{\sigma} \left(\sum_{\ell \in \mathbb{N}_d} \left| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i (x_i^k - x_{\lfloor \frac{n}{2} \rfloor + i}^k) (x_i^\ell - x_{\lfloor \frac{n}{2} \rfloor + i}^\ell) \right|^{2q'} \right)^{\frac{1}{q'}} \\
 &\leq \left(\sum_{\ell \in \mathbb{N}_d} \mathbb{E}_{\sigma} \left| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i (x_i^k - x_{\lfloor \frac{n}{2} \rfloor + i}^k) (x_i^\ell - x_{\lfloor \frac{n}{2} \rfloor + i}^\ell) \right|^{2q'} \right)^{\frac{1}{q'}}.
 \end{aligned}$$

Applying the Khinchin-Kahane inequality (Lemma 8 in the Appendix) with $q = 2q' = 4 \log d$ and $p = 2$ to the above inequality yields that

$$\begin{aligned}
 &\mathbb{E}_{\sigma} \sup_{\ell \in \mathbb{N}_d} \left| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i (x_i^k - x_{\lfloor \frac{n}{2} \rfloor + i}^k) (x_i^\ell - x_{\lfloor \frac{n}{2} \rfloor + i}^\ell) \right|^2 \\
 &\leq \left(\sum_{\ell \in \mathbb{N}_d} (2q')^{q'} \left[\mathbb{E}_{\sigma} \left| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i (x_i^k - x_{\lfloor \frac{n}{2} \rfloor + i}^k) (x_i^\ell - x_{\lfloor \frac{n}{2} \rfloor + i}^\ell) \right|^2 \right]^{q'} \right)^{\frac{1}{q'}} \\
 &= \left(\sum_{\ell \in \mathbb{N}_d} (2q')^{q'} \left[\sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} (x_i^k - x_{\lfloor \frac{n}{2} \rfloor + i}^k)^2 (x_i^\ell - x_{\lfloor \frac{n}{2} \rfloor + i}^\ell)^2 \right]^{q'} \right)^{\frac{1}{q'}} \\
 &\leq 2q' \sup_{x, x' \in \mathcal{X}} \|x - x'\|_{\infty}^2 d^{\frac{1}{q'}} \left[\sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} (x_i^k - x_{\lfloor \frac{n}{2} \rfloor + i}^k)^2 \right] \\
 &\leq 4e(\log d) \sup_{x, x' \in \mathcal{X}} \|x - x'\|_{\infty}^2 \left[\sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} (x_i^k - x_{\lfloor \frac{n}{2} \rfloor + i}^k)^2 \right]
 \end{aligned}$$

Putting the above estimation back into (26) implies that

$$\begin{aligned}
 R_n &\leq \sqrt{4e \log d} \left[\sup_{x, x' \in \mathcal{X}} \|x - x'\|_{\infty} \right] \mathbb{E}_{\mathbf{z}} \left(\sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \|x_i - x_{\lfloor \frac{n}{2} \rfloor + i}\|_F^2 \right)^{\frac{1}{2}} / \lfloor \frac{n}{2} \rfloor \\
 &\leq \sqrt{4e \log d} \left[\sup_{x, x' \in \mathcal{X}} \|x - x'\|_{\infty} \right] \left[\sup_{x, x' \in \mathcal{X}} \|x - x'\|_F \right] / \sqrt{\lfloor \frac{n}{2} \rfloor} \\
 &\leq 4\sqrt{e \log d} \left[\sup_{x, x' \in \mathcal{X}} \|x - x'\|_{\infty} \right] \left[\sup_{x, x' \in \mathcal{X}} \|x - x'\|_F \right] / \sqrt{n}.
 \end{aligned}$$

Combining this with Theorem 2 implies the inequality (25). This completes the proof of the example. \blacksquare

In the Frobenius-norm case, the main term of the bound (21) is $\mathcal{O}\left(\frac{\sup_{x, x' \in \mathcal{X}} \|x - x'\|_F^2}{\sqrt{n\lambda}}\right)$. This bound is consistent with that given by Jin et al. (2009) where $\sup_{x \in \mathcal{X}} \|x\|_F$ is assumed to be bounded by some constant B . Comparing the generalization bounds in the above examples. The key terms X_* and R_n mainly differ in two quantities, i.e. $\sup_{x, x' \in \mathcal{X}} \|x - x'\|_F$ and $\sup_{x, x' \in \mathcal{X}} \|x - x'\|_{\infty}$. We argue that $\sup_{x, x' \in \mathcal{X}} \|x - x'\|_{\infty}$ can be much less than $\sup_{x, x' \in \mathcal{X}} \|x - x'\|_F$. For instance, consider the input space $\mathcal{X} = [0, 1]^d$. It is easy to see that $\sup_{x, x' \in \mathcal{X}} \|x - x'\|_F = \sqrt{d}$ while $\sup_{x, x' \in \mathcal{X}} \|x - x'\|_{\infty} \equiv 1$. Consequently, we can summarise the estimations as follows:

- **Frobenius-norm:** In this case, we have $X_* = d$, and $R_n \leq \frac{2d}{\sqrt{n}}$.
- **Sparse L^1 -norm:** In this setting, we can see that $X_* = 1$, and $R_n \leq \frac{4\sqrt{e \log d}}{\sqrt{n}}$.
- **Mixed $(2, 1)$ -norm:** We obtain that $X_* = \sqrt{d}$, and $R_n \leq \frac{4\sqrt{ed \log d}}{\sqrt{n}}$.

Therefore, when d is large, the generalization bound with sparse L^1 -norm regularisation is much better than that with Frobenius-norm regularisation while the bound with mixed $(2, 1)$ -norm is between those with Frobenius norm and L^1 -norm. These theoretical results are nicely consistent with the rationale that sparse methods are more effective in dealing with high-dimensional data.

We end this section with two remarks. Firstly, in the setting of trace-norm regularisation, it remains a question to us on how to establish more accurate estimation of R_n by using the Khinchin-Kahane inequality. Secondly, the bounds in the above examples are true for similarity learning with different matrix-norm regularisation. Indeed, the generalization bound for similarity learning in Theorem 3 tells us that it suffices to estimate \tilde{X}_* and \tilde{R}_n . In analogy to the arguments in the above examples, we can get the following results. For similarity learning formulation (4) with Frobenius-norm regularisation, there holds

$$\tilde{X}_* = \sup_{x \in \mathcal{X}} \|x\|_F^2, \quad \tilde{R}_n \leq \frac{2 \sup_x \|x\|_F^2}{\sqrt{n}}.$$

For L^1 -norm regularisation, we have

$$\tilde{X}_* = \sup_{x \in \mathcal{X}} \|x\|_\infty^2, \quad \tilde{R}_n \leq 4 \sup_{x \in \mathcal{X}} \|x\|_\infty^2 \sqrt{e \log d} / \sqrt{n}.$$

In the setting of $(2, 1)$ -norm, we obtain

$$\tilde{X}_* = \sup_{x \in \mathcal{X}} \|x\|_\infty \sup_{x \in \mathcal{X}} \|x\|_F, \quad \tilde{R}_n \leq 4 \left[\sup_{x \in \mathcal{X}} \|x\|_F \sup_{x \in \mathcal{X}} \|x\|_\infty \right] \sqrt{e \log d} / \sqrt{n}.$$

Putting these estimations back into Theorem 3 yields generalization bounds for similarity learning with different matrix norms. For simplicity, we omit the details here.

5. Conclusion

In this paper we are mainly concerned with theoretical generalization analysis of the regularized metric and similarity learning. In particular, we first showed that the generalization analysis for metric/similarity learning reduces to the estimation of the Rademacher average over “sums-of-i.i.d.” sample-blocks. Then, we derived their generalization bounds with different matrix regularisation terms. Our analysis indicates that sparse metric/similarity learning with L^1 -norm regularisation could lead significantly better bounds than that with the Frobenius norm regularisation, especially when the dimension of the input data is high. Our novel generalization analysis develops the techniques of U-statistics (Peña and Giné, 1999; Clémençon et al., 2008) and Rademacher complexity analysis (Bartlett and Mendelson, 2002; Koltchinskii and Panchenko, 2002). In future we are planning to improve the generalization bounds for metric and similarity learning with trace-norm regularisation. The

target of supervised metric learning is to improve the generalization performance of kNN classifiers. It would be very interesting to investigate how the generalization performance of kNN classifiers relates to the generalization bounds of metric learning given here.

Acknowledgments

This work is supported by the EPSRC under grant EP/J001384/1. The corresponding author is Yiming Ying.

References

- A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning a mahalanobis metric from equivalence constraints. *J. of Machine Learning Research*, **6**: 937-965, 2005.
- P.L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *J. of Machine Learning Research*, **3**: 463-482, 2002.
- O. Bousquet and A. Elisseeff. Stability and generalization. *J. of Machine Learning Research* **2**: 499-526, 2002.
- D.R. Chen, Q. Wu, Y. Ying and D.X. Zhou. Support vector machine soft margin classifiers: error analysis, *J. of Machine Learning Research*, **5**: 1143-1175 (2004).
- G. Chechik, V. Sharma, U. Shalit, and S. Bengio. Large scale online learning of image similarity through ranking. *J. of Machine Learning Research*, **11**: 1109 -1135, 2010.
- S. Cl  mencon, G. Lugosi, and N. Vayatis. Ranking and empirical minimization of U-statistics. *Annals of Statistics*, **36**: 844-874, 2008.
- J. Davis, B. Kulis, P. Jain, S. Sra, and I. Dhillon. Information-theoretic metric learning. *ICML*, 2007.
- A. Globerson and S. Roweis. Metric learning by collapsing classes. *NIPS*, 2005.
- J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighbourhood component analysis. *NIPS*, 2004.
- M. Guillaumin, J. Verbeek and C. Schmid. Is that you? Metric learning approaches for face identification, *ICCV* 2009.
- S. C. H. Hoi, W. Liu, M. R. Lyu, and W.-Y. Ma. Learning distance metrics with contextual constraints for image retrieval. *CVPR*, 2006.
- R. Jin, S. Wang and Y. Zhou. Regularized distance metric learning: theory and algorithm, *NIPS*, 2009
- V. Koltchinskii and V. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, **30**, 1-5, 2002.
- P. Kar and P. Jain. Similarity-based learning via data-driven embeddings. *NIPS*, 2011.

- M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer Press, New York, 1991.
- A. Maurer. Learning similarity with operator-valued large-margin classifiers, *J. of Machine Learning Research*, **9**: 1049-1082, 2008.
- C. McDiarmid. Surveys in Combinatorics, Chapter On the methods of bounded differences, 148-188, 1989. Cambridge University Press, Cambridge (UK).
- V.H. De La Peña and E. Giné. *Decoupling: from Dependence to Independence*. Springer, New York, 1999.
- R. Rosales and G. Fung. Learning sparse metrics via linear programming, *KDD*, 2006.
- O. Shalit, D. Weinshall and G. Chechik. Online learning in the manifold of low-rank matrices. *NIPS*, 2010.
- C. Shen, J. Kim, L. Wang and A. Hengel. Positive semidefinite metric learning with boosting. *NIPS*, 2009.
- L. Torresani and K. Lee. Large margin component analysis. *NIPS*, 2007.
- K. Q. Weinberger and L. K. Saul. Fast solvers and efficient implementations for distance metric learning. *ICML*, 2008.
- E. Xing, A. Ng, M. Jordan, and S. Russell. Distance metric learning with application to clustering with side information. *NIPS*, 2002.
- L. Yang and R. Jin. Distance metric learning: A comprehensive survey. In *Technical report, Department of Computer Science and Engineering, Michigan State University*, 2007.
- Y. Ying, K. Huang and C. Campbell. Sparse metric learning via smooth optimisation. *NIPS*, 2009.
- Y. Ying and P. Li. Distance metric learning with eigenvalue optimisation. *J. of Machine Learning Research*, **13**: 1-26, 2012.

Appendix

In this appendix we assemble some facts, which were used to establish generalization bounds for metric/similarity learning.

Definition 4 We say the function $f : \prod_{k=1}^n \Omega_k \rightarrow \mathbb{R}$ with bounded differences $\{c_k\}_{k=1}^n$ if, for all $1 \leq k \leq n$,

$$\max_{z_1, \dots, z_k, z'_k, \dots, z_n} |f(z_1, \dots, z_{k-1}, z_k, z_{k+1}, \dots, z_n) - f(z_1, \dots, z_{k-1}, z'_k, z_{k+1}, \dots, z_n)| \leq c_k$$

Lemma 5 (*McDiarmid's inequality* ([McDiarmid, 1989](#))) Suppose $f : \prod_{k=1}^n \Omega_k \rightarrow \mathbb{R}$ with bounded differences $\{c_k\}_{k=1}^n$ then , for all $\epsilon > 0$, there holds

$$\Pr_{\mathbf{z}} \left\{ f(\mathbf{z}) - \mathbb{E}_{\mathbf{z}} f(\mathbf{z}) \geq \epsilon \right\} \leq e^{-\frac{2\epsilon^2}{\sum_{k=1}^n c_k^2}}.$$

Finally we list a useful property for U-statistics. Given the i.i.d. random variables $z_1, z_2, \dots, z_n \in \mathcal{Z}$, let $q : Z \times Z \rightarrow \mathbb{R}$ be a symmetric real-valued function. Denote a U-statistics of order two by $U_n = \frac{1}{n(n-1)} \sum_{i \neq j} q(x_i, x_j)$. Then, the U-statistic U_n can be expressed as

$$U_n = \frac{1}{n!} \sum_{\pi} \frac{1}{\lfloor \frac{n}{2} \rfloor} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} q(z_{\pi(i)}, z_{\pi(\lfloor \frac{n}{2} \rfloor + i)}) \quad (27)$$

where the sum is taken over all permutations π of $\{1, 2, \dots, n\}$. The main idea underlying this representation is to reduce the analysis to the ordinary case of i.i.d. random variable blocks. Based on the above representation, we can prove the following lemma which plays a critical role in deriving generalization bounds for metric learning. For completeness, we include a proof here. For more details on U-statistics, one is referred to [Cl  mencon et al. \(2008\)](#); [Pe  a and Gin   \(1999\)](#).

Lemma 6 Let $q_{\tau} : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$ be real-valued functions indexed by $\tau \in \mathcal{T}$ where \mathcal{T} is some index set. If z_1, \dots, z_n are i.i.d. then we have that

$$\mathbb{E} \left[\sup_{\tau \in \mathcal{T}} \frac{1}{n(n-1)} \sum_{i \neq j} q_{\tau}(z_i, z_j) \right] \leq \mathbb{E} \left[\sup_{\tau \in \mathcal{T}} \frac{1}{\lfloor \frac{n}{2} \rfloor} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} q_{\tau}(z_i, z_{\lfloor \frac{n}{2} \rfloor + i}) \right].$$

Proof From the representation of U-statistics (27), we observe that

$$\begin{aligned} \mathbb{E} \left[\sup_{\tau \in \mathcal{T}} \frac{1}{n(n-1)} \sum_{i \neq j} q_{\tau}(z_i, z_j) \right] &= \mathbb{E} \sup_{\tau} \frac{1}{n!} \sum_{\pi} \frac{1}{\lfloor \frac{n}{2} \rfloor} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} q_{\tau}(z_{\pi(i)}, z_{\pi(\lfloor \frac{n}{2} \rfloor + i)}) \\ &\leq \frac{1}{n!} \mathbb{E} \sum_{\pi} \sup_{\tau} \frac{1}{\lfloor \frac{n}{2} \rfloor} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} q_{\tau}(z_{\pi(i)}, z_{\pi(\lfloor \frac{n}{2} \rfloor + i)}) \\ &= \frac{1}{n!} \sum_{\pi} \mathbb{E} \sup_{\tau} \frac{1}{\lfloor \frac{n}{2} \rfloor} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} q_{\tau}(z_{\pi(i)}, z_{\pi(\lfloor \frac{n}{2} \rfloor + i)}) \\ &= \mathbb{E} \left[\sup_{\tau \in \mathcal{T}} \frac{1}{\lfloor \frac{n}{2} \rfloor} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} q_{\tau}(z_i, z_{\lfloor \frac{n}{2} \rfloor + i}) \right]. \end{aligned}$$

This completes the proof of the lemma. ■

We need the following contraction property of the Rademacher averages which is essentially implied by Theorem 4.12 in Ledoux and Talagrand [Ledoux and Talagrand \(1991\)](#), see also [Bartlett and Mendelson \(2002\)](#); [Koltchinskii and Panchenko \(2002\)](#).

Lemma 7 *Let F be a class of uniformly bounded real-valued functions on (Ω, μ) and $m \in \mathbb{N}$. If for each $i \in \{1, \dots, m\}$, $\Psi_i : \mathbb{R} \rightarrow \mathbb{R}$ is a function with $\Psi_i(0) = 0$ having a Lipschitz constant c_i , then for any $\{x_i\}_{i=1}^m$,*

$$\mathbb{E}_\epsilon \left(\sup_{f \in F} \left| \sum_{i=1}^m \epsilon_i \Psi_i(f(x_i)) \right| \right) \leq 2 \mathbb{E}_\epsilon \left(\sup_{f \in F} \left| \sum_{i=1}^m c_i \epsilon_i f(x_i) \right| \right). \quad (28)$$

The last property of Rademacher averages is the Khinchin-Kahane inequality (see e.g. [Peña and Giné \(1999, Theorem 1.3.1\)](#)).

Lemma 8 *For $n \in \mathbb{N}$, let $\{f_i \in \mathbb{R} : i \in \mathbb{N}_n\}$, and $\{\sigma_i : i \in \mathbb{N}_n\}$ be a family of i.i.d. Rademacher variables. Then, for any $1 < p < q < \infty$ we have*

$$\left(\mathbb{E}_\sigma \left| \sum_{i \in \mathbb{N}_n} \sigma_i f_i \right|^q \right)^{\frac{1}{q}} \leq \left(\frac{q-1}{p-1} \right)^{\frac{1}{2}} \left(\mathbb{E}_\sigma \left| \sum_{i \in \mathbb{N}_n} \sigma_i f_i \right|^p \right)^{\frac{1}{p}}$$